

# Hypothesis Testing For Network Data in Functional Neuroimaging

Cedric E. Ginestet\*, Prakash Balanchandran,  
Steven Rosenberg, and Eric D. Kolaczyk\*

*Department of Mathematics and Statistics  
Boston University, Boston, MA.*

**Abstract:** In recent years, it has become common practice in neuroscience to use networks to summarize relational information in a set of measurements, typically assumed to be reflective of either functional or structural relationships between regions of interest in the brain. One of the most basic tasks of interest in the analysis of such data is the testing of hypotheses, in answer to questions such as “Is there a difference between the networks of these two groups of subjects?” In the classical setting, where the unit of interest is a scalar or a vector, such questions are answered through the use of familiar two-sample testing strategies. Networks, however, are not Euclidean objects, and hence classical methods do not directly apply. We address this challenge by drawing on concepts and techniques from geometry, and high-dimensional statistical inference. Our work is based on a precise geometric characterization of the space of graph Laplacian matrices and a nonparametric notion of averaging due to Fréchet. We motivate and illustrate our resulting methodologies for testing in the context of networks derived from functional neuroimaging data on human subjects from the 1000 Functional Connectomes Project. In particular, we show that this global test is more statistically powerful, than a mass-univariate approach.

**AMS 2000 subject classifications:** Fréchet mean, fMRI, Graph Laplacian, Hypothesis Testing, Matrix manifold, Network data, Neuroscience.

## 1. Introduction

Functional neuroimaging data has been central to the advancement of our understanding of the human brain. Neuroimaging data sets are increasingly approached from a graph-theoretical perspective, using the tools of modern network science (Bullmore and Sporns, 2009). This has elicited the interest of statisticians working in that area. At the level of basic measurements, neuroimaging data can be said to consist typically of a set of signals (usually time series) at each of a collection of pixels (in two dimensions) or voxels (in three dimensions). Building from such data, various forms of higher-level data representations are employed in neuroimaging. Traditionally, two- and three-dimensional

---

\*This work was supported by a grant from the Air Force Office for Scientific Research (AFOSR), whose grant number is FA9550-12-1-0102. The data from the 1000 Functional Connectome Project was accessed through the International Neuroimaging Data-sharing Initiative (INDI), which was designed for unrestricted data-sharing via the Neuroimaging Informatics Tool and Resources Clearinghouse (NITRC). We are also indebted to Sean Markan, Lizhen Lin, Emily Stephen and Heather Shappell for useful suggestions and discussion.

images have, naturally, been the norm, but increasingly in recent years there has emerged a substantial interest in network-based representations.

### 1.1. *Motivation*

Let  $G = (V, E)$  denote a graph, based on  $d = |V|$  vertices. In this setting, the vertices  $v \in V$  correspond to regions of interest (ROIs) in the brain, often pre-defined through considerations of the underlying neurobiology (e.g., the putamen or the cuneus). Edges  $\{u, v\} \in E$  between vertices  $u$  and  $v$  are used to denote a measure of association between the corresponding ROIs. Depending on the imaging modality used, the notion of ‘association’ may vary. For example, in diffusion tensor imaging (DTI), associations are taken to be representative of structural connectivity between brain regions. On the other hand, in functional magnetic resonance imaging (fMRI), associations are instead thought to represent functional connectivity, in the sense that the two regions of the brain participate together in the achievement of some higher-order function, often in the context of performing some task (e.g., counting from 1 to 10).

With neuroimaging now a standard tool in clinical neuroscience, and with the advent of several major neuroscience research initiatives – perhaps most prominent being the recently announced Brain Research Accelerated by Innovative Neurotechnologies (BRAIN) initiative – we are quickly moving towards a time in which we will have available databases composed of large collections of secondary data in the form of network-based data objects. Faced with databases in which networks are a fundamental unit of data, it will be necessary to have in place the statistical tools to answer such questions as, “What is the ‘average’ of a collection of networks?” and “Do these networks differ, on average, from a given nominal network?,” as well as “Do two collections of networks differ on average?” and “What factors (e.g., age, gender, etc.) appear to contribute to differences in networks?”, or finally, say, “Has there been a change in the networks for a given subpopulation from yesterday to today?” In order to answer these and similar questions, we require network-based analogues of classical tools for statistical estimation and hypothesis testing.

While these classical tools are among the most fundamental and ubiquitous in use in practice, their extension to network-based datasets, however, is not immediate and, in fact, can be expected to be highly non-trivial. The main challenge in such an extension is due to the simple fact that networks are not Euclidean objects (for which classical methods were developed) – rather, they are combinatorial objects, defined simply through their sets of vertices and edges. Nevertheless, our work here in this paper demonstrates that networks can be associated with certain natural subsets of Euclidean space, and furthermore demonstrates that through a combination of tools from geometry, probability on manifolds, and high-dimensional statistical analysis it is possible to develop a principled and practical framework in analogy to classical tools. In particular, we focus on the development of an asymptotic framework for one- and two-sample hypothesis testing.

Key to our approach is the correspondence between an undirected graph  $G$  and its Laplacian, where the latter is defined as the matrix  $L = D - W$ ; with  $W$  denoting the  $d \times d$  adjacency matrix of  $G$  and  $D$  a diagonal matrix with the vertex degrees along the diagonal. When  $G$  has no self-loops and no multi-edges, the correspondence between graphs  $G$  and Laplacians  $L$  is one-to-one. Our work takes place in the space of graph Laplacians. Importantly, this work requires working not in standard Euclidean space  $\mathbb{R}^n$ , but rather on certain subsets of Euclidean space which are either submanifolds of  $\mathbb{R}^n$  or submanifolds with corners of  $\mathbb{R}^n$ . While these subsets of Euclidean space have the potential to be complicated in nature, we show that in the absence of any nontrivial structural constraints on the graphs  $G$ , the geometry of these subsets is sufficiently ‘nice’ to allow for a straightforward definition of distance between networks to emerge.

Our goal in this work is the development of one- and two-sample tests for network data objects that rely on a certain sense of ‘average’. We adopt the concept of Fréchet means in defining what average signifies in our context. Recall that, for a metric space,  $(\mathcal{X}, \rho)$ , and a probability measure,  $Q$ , on its Borel  $\sigma$ -field, under appropriate conditions, the Fréchet mean of  $Q$  is defined as the (possibly nonunique) minimizer

$$\mu := \operatorname{argmin}_{x \in \mathcal{X}} \int_{\mathcal{X}} \rho^2(x, y) Q(dy). \quad (1)$$

Similarly, for any sample of realizations from  $Q$  on  $\mathcal{X}$ , denoted  $Y := \{Y_1, \dots, Y_n\}$ , the corresponding sample Fréchet mean is defined as

$$\hat{\mu}_n(Y) := \operatorname{argmin}_{x \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^n \rho^2(x, Y_i). \quad (2)$$

Thus, the distance  $\rho$  that emerges from our study of the geometry of the space of networks implicitly defines a corresponding notion of how to ‘average’ networks.

Drawing on results from nonparametric statistical inference on manifolds, we are then able to establish a central limit theory for such averages and, in turn, construct the asymptotic distributions of natural analogues of one- and two-sample  $z$ -tests. These tests require knowledge of the covariance among the edges of our networks, which can be expected to be unavailable in practice. Nevertheless, we show how recent advances in the estimation of large, structured covariance matrices can be fruitfully brought to bear in our context, and provide researchers with greater statistical power than a mass-univariate approach, which is the standard approach in this field.

### ***1.2. The 1000 Functional Connectomes Project***

Our approach is motivated by and illustrated with data from the 1000 Functional Connectomes Project (FCP). This major MRI data-sharing initiative was launched in 2010 (Biswal et al., 2010). The impetus for the 1000 FCP was given

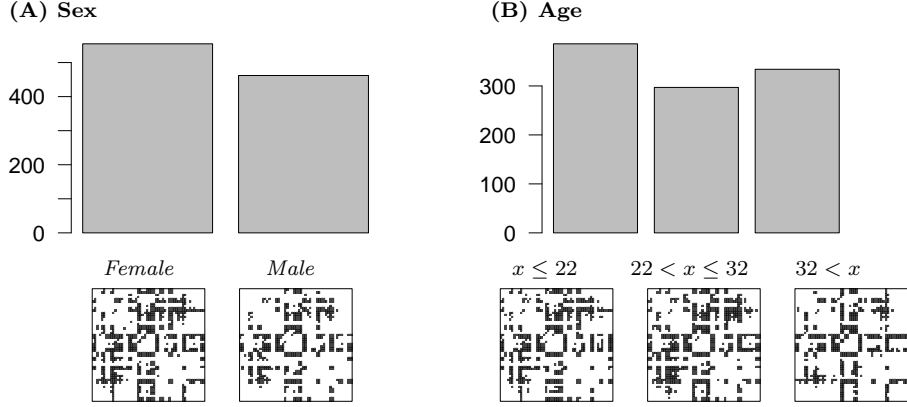
by a need to make widely accessible neuroimaging data, which are costly and time-consuming to collect (Biswal et al., 2010). This was conducted within the so-called “discovery science” paradigm, paralleling similar initiatives in systems biology. The 1000 FCP constituted the largest data set of its kind, at the time of its release. As for the use of such large data sets in genetics, it is believed that facilitating access to high-throughput data generates economies of scale that are likely to lead to more numerous and more substantive research findings.

The 1000 FCP describes functional neuroimaging data from 1093 subjects, located in 24 community-based centers. The mean age of the participants is 29 years, and all subjects were 18 years-old or older. Each individual scan lasted between 2.2 and 20 minutes. The strength of the MRI scanner varied across centers, with  $n = 970$  scans at 3T and  $n = 123$  at 1.5T. Voxel-size was 1.5–5mm within the plane; and slice thickness was 3–8mm. The ethics committee in each contributing data center approved the project; and the institutional review boards of the NYU Langone Medical Center and of the New Jersey Medical School approved the dissemination of the data. This freely available data set has been extensively used in the neuroimaging literature (Yan et al., 2013; Tomasi and Volkow, 2010; Zuo et al., 2012).

The individual fMRI scans were parcellated into a set of 50 cortical and subcortical regions, using the Automated Anatomical Labeling (AAL) template (Tzourio-Mazoyer et al., 2002). The voxel-specific time series in each of these regions were aggregated to form mean regional time series, as commonly done in the study of the human connectome (see for example Achard et al., 2006). The resulting regional time series were then compared using two different measures of association. We here considered the correlation coefficient since this measure has proved to be popular in the neuroimaging literature (Ginestet and Simmons, 2011; Pachou et al., 2008; Micheloyannis et al., 2009).

Subjects in the 1000 FCP data can be subdivided with respect to sex. Several groups of researchers have previously considered the impact of sex differences on resting-state connectivity (Biswal et al., 2010; Tomasi and Volkow, 2011). It is hypothesized that sexual dimorphism in human genomic expression is likely to affect a wide range of physiological variables (Ellegren and Parsch, 2007). In particular, differences in hormonal profiles (e.g. estrogen) during brain development is known to be related to region-specific effects (McEwen, 1999). Thus, it is of interest to compare the subject-specific networks of males and females in the 1000 FCP data set. Observe that previous research in this field has established *local* sex differences in connectivity by considering individual edge weights (Biswal et al., 2010; Tomasi and Volkow, 2011). By contrast, we are here investigating the effect of sex differences on *entire* networks.

It is here useful to distinguish between these two types of network data analysis in neuroimaging. While local analysis focuses on edge-specific statistics; global analysis instead considers network topological properties such as the shortest-path length. In this paper, we are extending the latter by providing a framework for identifying the mean network, and characterizing the space of all possible such networks. In the sequel, we will also be interested in evaluating age differences, as well as collection-site differences in network connectivity.



**Fig 1.** Descriptive statistics for the 1000 FCP data set. In panel (A), the proportions of males and females in the data set is provided with the corresponding group-specific mean Laplacians for networks over 50 AAL vertices. Similarly, in panel (B), the age variable has been divided into three groups, and the respective means are reported for each age group. The Laplacians have been binarized with respect to the 75<sup>th</sup> percentile of the overall distribution of entries in the full 1000 FCP database. (Black indicates entries greater or equal than that percentile).

The organization of this paper is as follows. In Section 2, we describe the statistical and mathematical background of this type of research questions. In Section 3, we provide a geometrical characterization of the space of networks under scrutiny. In Section 4, we describe how certain central limit theorems can be adapted to this space, in order to construct a statistical inferential framework for network data. A simulation study exploring the relationship between statistical power and various aspects of neuroimaging data is reported in Section 5. In Section 6, we apply this framework to the analysis of a subset of the data from the 1000 FCP. These results and the potential extensions of the proposed statistical tests are then discussed in Section 7.

## 2. Related Work

At the heart of the class of statistical problems we wish to address is a desire to summarize and compare groups of network data objects in a statistically principled manner. There are, of course, already a variety of numerical devices available for carrying out certain descriptive summaries and comparisons. Basic set-theoretic operations (e.g., union, intersection, symmetric difference) are all well-defined for graphs. More broadly, various metrics, such as the Hamming distance, have been borrowed from other fields and applied to graphs. Currently, the mainstay in the analysis of network data in neuroimaging, is the mass-univariate approach in which independent tests are conducted for every edge, adjusting for multiple testing. See Ginestet, Fournel and Simmons (2014) for a survey of such methods in the context of functional neuroimaging.

Such mass-univariate approaches, however, fail to draw inference about networks as a whole. In particular, it is unclear whether multiple local differences necessarily lead to globally significant differences. One may tackle this problem by treating network data objects as data points. What is lacking to achieve this, however, is the necessary mathematical foundation – establishing a formal ‘space’ of graphs, equipped with a formal metric, with understood geometric and topological properties, so that a formal notion of probability and measure can be defined, all underlying the desired theory and methods for the hypothesis testing problems of interest here.

Networks are not the only data type for which standard Euclidean-based methods are insufficient. Statistical inference on manifolds – in particular on spheres and shapes spaces – has a fairly long history. There is a substantial literature on statistics on spheres, or so-called directional statistics, going back to a seminal paper by R.A. Fisher in 1953 (Fisher, 1953), and works by Watson (1983), Mardia and Jupp (2009), and Fisher, Lewis and Embleton (1987), among others. Statistical analysis on shapes that are landmark-based was pioneered by Kendall (1977), Kendall (1984) and Bookstein (1978). Inference in these settings takes various forms. Nonparametric forms of inference typically employ a notion of averaging due to Fréchet (1948), as we do in this paper. Nevertheless, little work has been pursued with manifolds given as some general metric space – such as the spaces of networks that are our main interest. The most related work seems to be due to Billera, Holmes and Vogtmann (2001) and Barden, Le and Owen (2013), who study the metric geometry of the space of phylogenetic trees and derive a central limit theorem for the Fréchet mean in such spaces. Also see the related work of Marron and colleagues in the context of so-called object-oriented data analysis with trees (Wang and Marron, 2007; Aydin et al., 2009).

In order to establish a formal characterization of a well-defined ‘space’ of networks, it is natural to associate a network with a matrix. And, while there are several such matrices that might be used, we have found that the (combinatorial) graph Laplacian is particularly appropriate. The Laplacian falls in the cone of symmetric positive (semi)definite (PSD) matrices. A substantial amount of effort has been expended on uncovering the mathematical properties of the PSD cone (Bhatia, 1997; Moakher and Zerai, 2011). In addition, there has in recent years been quite a lot of work exploring the various notions of ‘average’ induced upon this manifold by the underlying choices of geometry (Arsigny et al., 2007; Moakher, 2005; Bonnabel and Sepulchre, 2009). Finally, depending on the choice of average adopted, there are results establishing the probabilistic and statistical properties of averages through CLTs (Bhattacharya and Patrangenaru, 2003, 2005; Bhattacharya and Bhattacharya, 2012; Kendall and Le, 2011). Much of this research has been motivated by shape analysis (Le and Kume, 2000; Le, 2001), but many of these results have been developed in other areas of applications where matrices play a key role such as in DTI (Dryden, Koloydenko and Zhou, 2009).

However, the space of graph Laplacians forms a *subset* of the PSD cone and, furthermore, by definition this subset intersects in a non-trivial fashion with the

boundary of this cone. Therefore, results for PSD matrices do not carry over immediately to the space of graph Laplacians – the latter must necessarily be studied in its own right. At present, while graph Laplacians as individual objects are well-studied –see Chung (1997), who discusses discrete eigenvalue and isoperimetric estimates analogous to Riemannian estimates (see also Chavel, 1984; Xia, 2013) – there appears to be no formal body of results to date establishing the properties of the *space* of graph Laplacians – and certainly none that reflects the impact of what have become established canonical properties of complex networks (e.g., sparseness, small-world, etc.). The closest work of which we are aware is, for example, recent work in the signal processing literature, characterizing subspaces of the PSD cone corresponding to subsets of covariance matrices sharing certain simple structural properties such as rank or trace constraints (Krishnamachari and Varanasi, 2013).

A certain notion of embedding is crucial to the mathematical and probabilistic theory underlying our approach. There are, in fact, different uses of the term “embedding”. Our work involves averaging or comparing different networks/graphs via the distance between network Laplacians computed by first embedding (i.e. smoothly injecting) the set of Laplacian matrices into a Euclidean space; here “embedding” is defined as in the differentiable topology literature (see chap. 7 in Lee, 2006). This seems to have advantages over comparing networks via e.g. isometric embeddings of the graph itself into  $\mathbb{R}^3$ , for which computation of the types of distance functions that have been useful (e.g. Gromov-Hausdorff distance) is impractical.

In addition, there is also the large literature on graph embedding, which maps a graph onto a typically low-dimensional Euclidean space using eigenvector/eigenvalue information of the adjacency matrix or associated Laplacian (Linial, London and Rabinovich, 1995; Linial, 2002; Yan et al., 2007; Fu and Ma, 2013). Graph embedding methods are very different from differentiable topology techniques. In particular, the image of a graph embedding is often used as a dimension-reduction tool. This map in general has some distortion, and so is not an isometry. This change in the geometry from the domain space to the range space implies that the precise inference framework for manifolds that we employ here, as described below, cannot be applied to graph embeddings. Thus, there is no natural notion of average and projection onto the image under a graph embedding, and in fact such a projection may not exist. On the other hand, our notion of embedding, which considers the spaces of Laplacians as a manifold, does not reduce dimension, preserves all the raw information in a specific graph, and allows analysis of averages and projections by geometric methods.

### 3. Characterization of Spaces of Networks

In this section, we establish the necessary mathematical properties associated with a certain notion of a ‘space’ of networks, from which a natural notion of ‘averaging’ emerges. In fact, we offer several variations of a space of networks and, in doing so, illustrate how even relatively simple constraints on network

topology affect the geometry of these spaces. The geometry is important when seeking to develop the corresponding probabilistic behavior of averages of networks, as we do in section 4, which also informs the sampling distributions of the one- and two-sample test statistics that we develop.

### 3.1. Main Results

Let  $G = (V, E, W)$  be a *weighted* undirected graph, for weights  $w_{ij} = w_{ji} \geq 0$ , where equality with zero holds if and only if  $\{i, j\} \notin E$ . Assume  $G$  to be simple (i.e., no self-loops or multi-edges). We associate uniquely with each graph  $G$  its graph Laplacian  $L = D(W) - W$ , where  $D$  is a diagonal matrix of weighted degrees (also called vertex strengths), i.e.,  $D_{jj} = d_j(W) = \sum_{i \neq j} w_{ij}$ . We further assume in most of what follows that  $G$  is connected, in which case  $L$  has one (and only one) zero eigenvalue and all the others are positive (and hence  $L$  is positive semi-definite).

Under the assumption that  $G$  is simple, there is a one-to-one correspondence between graphs  $G$  and Laplacian matrices  $L$ . We therefore define our space of networks through a corresponding space of Laplacians. In the following theorem, we show that an initial notion of the space of graph Laplacians over  $d$  nodes admits a relatively simple topology, which can be described as a convex subset of an affine space in  $\mathbb{R}^{d^2}$ .

**Theorem 1.** *The set  $\mathcal{L}_d$  of  $d \times d$  matrices  $A$ , satisfying:*

- (1) *Rank( $A$ ) =  $d - 1$ ,*
- (2) *Symmetry,  $A' = A$ ,*
- (3) *Positive semi-definiteness,  $A \geq 0$ ,*
- (4) *The entries in each row sum to 0,*
- (5) *The off-diagonal entries are negative,  $a_{ij} < 0$ ;*

*forms a submanifold of  $\mathbb{R}^{d^2}$  of dimension  $d(d - 1)/2$ . In fact,  $\mathcal{L}_d$  is a convex subset of an affine space in  $\mathbb{R}^{d^2}$  of dimension  $d(d - 1)/2$ .*

A proof of this theorem is in Appendix A. The practical importance of this result is that  $\mathcal{L}_d$  admits (many) Riemannian metrics, which give rise to a restricted class of distance functions. For example, any one of these metrics turns  $\mathcal{L}_d$  into a length space in the sense of Gromov (2001), i.e. the distance between any two points  $A, B \in \mathcal{L}_d$  is the length of some path from  $A$  to  $B$ . Also, all the usual notions of curvature, and its influence on variations of geodesics, come into play.

However, we note that the definition of  $\mathcal{L}_d$  requires that *every* potential edge in  $G$  be present, with edges distinguished purely by the relative magnitude of their weights. Consider the description of the 1000 FCP data in Section 1.2. For the case where our network is defined to be, say, the matrix  $W$  of empirical correlations or mutual information of signals between pairs of ROIs, the space  $\mathcal{L}_d$  is appropriate. On the other hand, if we chose instead to work with a thresholded version of such matrices, then it is important that we allow for *both* the



presence/absence of edges by allowing weights to be zero. The result of Theorem 1 can be extended to include such networks, as described in the following corollary. This leads to a manifold that possesses corners. A good introduction to manifolds with corners can be found in standard texts on smooth manifolds (see chap. 14 in Lee, 2006). Moreover, this manifold is also a convex subset of Euclidean space.

**Corollary 1.** *In Theorem 1, if condition (5) is replaced by*

*(5') The off-diagonal entries are non-positive,  $a_{ij} \leq 0$ ;*

*then the corresponding matrix space  $\mathcal{L}'_d$  is a manifold with corners of dimension  $d(d-1)/2$ . Furthermore,  $\mathcal{L}'_d$  is a convex subset of an affine space in  $\mathbb{R}^{d^2}$  of dimension  $d(d-1)/2$ .*

A proof of this corollary is provided in Appendix A. Importantly, the above theorem and its corollary indicate that the Euclidean metric (i.e. the Frobenius distance on the space of  $d \times d$  matrices with real-valued entries) is a natural choice of distance function on our spaces of Laplacians. The metric space of interest is therefore composed of, for example,  $(\mathcal{L}'_d, \rho_F)$ , where  $\rho_F$  is the Frobenius distance

$$\rho_F(X, Y) := \|X - Y\|_F^2 = \sum_{i,j}^d (x_{ij} - y_{ij})^2 \quad ,$$

for any pair of matrices  $X, Y \in \mathcal{L}'_d$ . As we shall see momentarily below, in Section 4, the concept of a Fréchet mean and its sample-based analogue, as detailed in equations (1) and (2), may now be brought to bear, yielding a well-defined sense of an average of networks.

### 3.2. Extensions: Implications of constraints on network topology

In ending this section, we note that our definition of a ‘space of networks’ is intentionally minimal in lacking constraints on the topology of the networks. However, one of the most fundamental results that has emerged from the past 20 years of complex network research is the understanding that real-world networks typically (although not exclusively) tend to possess a handful of quite marked structural characteristics. Examples include sparseness (i.e., number of edges scaling like the number of vertices), heavy-tailed degree distributions, and the presence of cohesive subgraphs (a.k.a. communities). See chap. 8 in Newman (2010), for example, for details and a more comprehensive summary. Importantly, this fact suggests that the appropriate differential or metric measure geometry of the ‘space of all networks’ – or, more formally, the space of Laplacians corresponding to such networks – depends on the constraints imposed on these networks/Laplacians.

While a detailed study of these implications are beyond the scope of this paper, we illustrate them through the following theorem, which extends the

previous results to the more general case of graphs composed of different numbers of connected components. In particular, we can generalize Theorem 1 to spaces of Laplacians representing graphs with a fixed number of components,  $\ell$ .

**Theorem 2.** *The set  $\mathcal{L}_\ell$  of  $d \times d$  matrices  $E$  satisfying*

- (1 $_\ell$ )  $\text{Rank}(A) = \ell$ ,
- (2)  $E$  is symmetric,
- (3)  $E$  is positive semidefinite,
- (4) The sum of the entries of each column is zero,
- (5) Each off-diagonal entry is negative;

*forms a submanifold of  $\mathbb{R}^{d^2}$  of dimension  $d\ell - \ell(\ell + 1)/2$ .*

A proof of this theorem is in Appendix A. Intuitively, this result is stating that the number of connected components of the average of two graphs can be smaller than the number of components of each graph, but it cannot be larger. That is, the average of two graphs may decrease the number of communities, but it cannot increase that number. Indeed, when taking the Euclidean average of several graphs with non-negative edge weights, we can only maintain existing edges or create new edges.

#### 4. Statistical Inference on Samples of Networks

Having characterized a space of networks, it becomes possible to construct an inferential framework for comparing one or more samples of networks. We here describe some analogues of the classical one- and two-sample  $t$ -statistics in this setting. These are obtained by first selecting a notion of averaging and deriving a central limit theorem for sequences of network averages, next appealing to Wald-like constructions of test statistics, and finally, utilizing recent results on high-dimensional covariance estimation.

##### 4.1. A Central Limit Theorem

Let  $G_1, \dots, G_n$  denote  $n$  graphs, each simple and assumed to have the same number of vertices  $d$ ; and let  $L_1, \dots, L_n$  be the corresponding combinatorial Laplacians. The  $L_i$ 's are assumed to be independent and identically distributed according to a distribution  $Q$ . In the context of neuroimaging, for example, these might be the correlation networks from resting-state fMRI images obtained from a group of human subjects matched for various demographic characteristics (e.g., age, gender) and health status (e.g., clinical manifestation of a given neurodegenerative disease).

The results of the previous section tell us that an appropriate sense of distance between pairs of networks is given by the Euclidean distance between their corresponding Laplacians. Combining these results with the definition of average in equations (1) and (2), indicates that a principled way in which to define the average of  $n$  networks is through elementwise averaging of the entries of their

Laplacians (and hence their adjacency matrices). Such an average is, of course, easily computed. However, this is not always the case when computing averages on manifolds. See, for instance, chap. 6 in Bhatia (2007) for an illustration of the difficulties that may arise, when computing the matrix mean in the cone of positive-definite symmetric matrices with respect to the geodesic distance on that manifold.

In the context of the 1000 FCP database, we wish to compare networks with respect to the sex of the subjects, over different age group, and over various collection sites. It is thus necessary to compute the means in each subgroup of networks. This was done, for example, in Figure 1, by constructing the Euclidean mean of the Laplacians for each group of subjects in different age groups. Such group-specific mean Laplacians can then be interpreted as the mean functional connectivity in each group.

The sample Fréchet mean  $\widehat{L}_n$  is a natural statistic upon which to build our hypothesis tests about the average of networks or groups of networks. In order to do so, we require an understanding of the behavior of  $\widehat{L}_n$  as a random variable. Under broad regularity conditions,  $\widehat{L}_n \rightarrow \Lambda$  almost surely; that is, the sample Fréchet mean,  $\widehat{L}_n$ , is a consistent estimator of the true mean  $\Lambda$  (see Ziezold, 1977). In addition, under further assumptions, we can also derive a central limit theorem for the sample Fréchet mean of Laplacians, with respect to the half-vectorization map,  $\phi$ .

**Theorem 3.** *If the expectation,  $\Lambda := \mathbb{E}[L]$ , does not lie on the boundary of  $\mathcal{L}'_d$ , and  $\mathbb{P}[U] > 0$ , where  $U$  is an open subset of  $\mathcal{L}'_d$  with  $\Lambda \in U$ , and under some further regularity conditions (see appendix B); we obtain the following convergence in distribution,*

$$n^{1/2}(\phi(\widehat{L}_n) - \phi(\Lambda)) \longrightarrow N(0, \Sigma),$$

where  $\Sigma := \mathbb{Cov}[\phi(L)]$  and  $\phi(\cdot)$  denotes the half-vectorization of its matrix argument.

A proof of this theorem and the full set of assumptions are provided in appendix B. The argument is a specialization of a general result due to Bhattacharya and Lin (2013). The result stated in the theorem has fundamental significance regarding our goal of developing analogues of classical testing strategies for the analysis of network data objects. It is an asymptotic result stating that, given a sufficient number of samples from a population of networks, an appropriately defined notion of sample average behaves in a classical manner: It possesses a statistical distribution that is approximately multivariate normal, centered on the population mean  $\mu$  and with covariance  $\Sigma$ .

#### 4.2. One-sample, Two-sample and k-sample Tests

As an immediate consequence of this central limit theorem, we can define natural analogues of classical one- and two-sample hypothesis tests. Consider, for example, the null hypothesis that the expectation  $\Lambda = \mathbb{E}[L]$  is equal to some

pre-specified value, i.e.,  $H_0 : \Lambda = \Lambda_0$ . In the context of neuroimaging, the choice of  $\Lambda_0$  might correspond to a reference connectivity pattern, derived from a large study, such as the 1000 FCP, for instance. In addition to the conditions stated in Theorem 3, let us now assume that the true covariance matrix,  $\Sigma$ , is *non-singular*. Moreover, it is also assumed that the target Laplacian,  $\Lambda_0$ , is known. Then, we are immediately led to a test statistic with an asymptotic  $\chi^2$ -distribution. (For expediency, we will now drop the subscript  $n$  in  $\hat{L}_n$ .)

**Corollary 2.** *Under the assumptions of Theorem 3, and under the null hypothesis  $H_0 : \mathbb{E}[L] = \Lambda_0$ , we have,*

$$T_1 := n(\phi(\hat{L}) - \phi(\Lambda_0))' \hat{\Sigma}^{-1} (\phi(\hat{L}) - \phi(\Lambda_0)) \longrightarrow \chi_m^2,$$

with  $m := \binom{d}{2}$  degrees of freedom, and where  $\hat{\Sigma}$  is the sample covariance.

Similarly, one can also construct a statistical test for two or more independent samples using the same framework. Assume that we have  $k$  independent sets of Laplacians of dimension  $d \times d$ , and consider the problem of testing whether or not these sets have in fact been drawn from the same population. Each sample of Laplacians has the form,  $L_{in_j}$ , where  $i = 1, \dots, n_j$ ; for every  $j = 1, \dots, k$ . Each of these  $k$  populations has an unknown mean, denoted  $\Lambda_j$ , while the sample means of these sets of Laplacians are denoted by  $\hat{L}_j$ , for each  $j = 1, \dots, k$ , respectively. Then, as a direct corollary to Theorem 3, we have the following asymptotic result.

**Corollary 3.** *Assume that every  $\Lambda_j$  does not lie on the boundary of  $\mathcal{L}'_d$ , and that  $\mathbb{P}[U] > 0$ , where  $U$  is an open subset of  $\mathcal{L}'_d$ , and where  $L_j \in U$ , for every  $j = 1, \dots, k$ . Moreover, also assume that  $n_j/n \rightarrow p_j$  for every sample, with  $n := \sum_j n_j$ , and  $0 < p_j < 1$ . Then, under  $H_0 : \Lambda_1 = \dots = \Lambda_k$ , we have*

$$T_k := \sum_{j=1}^k n_j (\phi(\hat{L}_j) - \phi(\hat{L}))' \hat{\Sigma}^{-1} (\phi(\hat{L}_j) - \phi(\hat{L})) \longrightarrow \chi_{(k-1)m}^2,$$

where  $\hat{L}_j$  denotes the sample mean of the  $j^{\text{th}}$  sample,  $\hat{L}$  represents the grand sample mean of the  $n$  Laplacians, and  $\hat{\Sigma} := \sum_{j=1}^k \hat{\Sigma}_j / n_j$  is a pooled estimate of covariance, with the  $\hat{\Sigma}_j$ 's denoting the individual sample covariance matrices of each subsample, with respect to  $\hat{L}$ .

Hence, we can compare the test statistic  $T_k$  against an asymptotic chi-square distribution in assessing the evidence against the null hypothesis stating that all  $k$  population means are identical –that is,  $H_0 : \Lambda_1 = \dots = \Lambda_k$ . As a special case of this corollary, we obtain the following two-sample test statistic, which evaluates whether the null hypothesis,  $H_0 : \Lambda_1 = \Lambda_2$ , is true:

$$T_2 := (\phi(\hat{L}_1) - \phi(\hat{L}_2))' \hat{\Sigma}^{-1} (\phi(\hat{L}_1) - \phi(\hat{L}_2)) \longrightarrow \chi_m^2,$$

where as before  $m := \binom{d}{2}$ , and the pooled sample covariance matrix is given by  $\hat{\Sigma} = \hat{\Sigma}_1 / n_1 + \hat{\Sigma}_2 / n_2$ .

### 4.3. Covariance Estimation

We note that in order to use any of the above results in a practical setting, we must have knowledge of the covariance matrix  $\Sigma = \text{Cov}[\phi(L)]$ . It can be expected that we must use a sample-based estimate. However, because the order of this matrix is  $O(d^2) \times O(d^2)$ , and the sample size  $n$  is potentially much smaller than  $O(d^2)$ , the traditional sample covariance  $\hat{\Sigma}$  is likely to be numerically unstable, and is not guaranteed to be positive definite.

Fortunately, the development of estimators of  $\Sigma$  in such low-sample/high-dimension contexts has been an active area of statistical research over the past few years. Typically, borrowing regularization strategies from the field of nonparametric function estimation, optimization of a cost function combining Frobenius norm or penalized maximum likelihood with a regularization term yields a convex optimization problem that can be solved efficiently. Generally, the choice of a regularization term is linked to the assumed structure of the covariance matrix – for example, assumptions of banding (Bickel and Levina, 2008b) or sparseness (Bickel and Levina, 2008a; Cai and Liu, 2011; Karoui, 2008). There is also a substantial recent literature on the closely related problem of estimating the inverse covariance matrix  $\Sigma^{-1}$ . See Cai, Liu and Luo (2011) for a recent example and associated citations.

In the context of neuroimaging, it can be expected that the networks of interest will be sparse (Lee et al., 2011). That is, it can be expected that the number of edges  $|E|$  present in a network  $G = (V, E)$  will be roughly on the same order of magnitude as the number of vertices  $d = |V|$ . Empirically, across the 1000 FCP data that is our focus in this paper, we have found that the covariance matrix,  $\Sigma$ , of the entries of the Laplacians of these functional networks also tended to be sparse, thereby justifying a sparse estimation procedure for  $\Sigma$ .

Accordingly, as an alternative to the sample covariance, we adopt the use of an estimator due to Cai and Liu (2011), which is a penalized maximum likelihood estimator under Gaussianity assumptions that possesses an optimal rate of convergence. We briefly describe this estimator here. For a generic sample  $X_1, \dots, X_n$  of independent and identically distributed random variables, define

$$\Sigma^* := \frac{n-1}{n} \hat{\Sigma} = [\sigma_{ij}^*]_{1 \leq i, j \leq d},$$

where  $\hat{\Sigma}$  is the sample covariance. This estimator can be thresholded in the following manner in order to obtain a new estimator of the population covariance matrix,  $\tilde{\Sigma} := [s_{\lambda_{ij}}(\sigma_{ij}^*)]_{1 \leq i, j \leq d}$ , where the thresholding function is defined as follows,

$$s_{\lambda_{ij}}(\sigma_{ij}^*) := \sigma_{ij}^* \mathcal{I}\{\sigma_{ij}^* \geq \lambda_{ij}\},$$

with  $\mathcal{I}\{\cdot\}$  denoting the indicator function. Moreover, the weights,  $\lambda_{ij}$ , are given by  $\lambda_{ij} := \delta(\hat{\theta}_{ij} \log(d)/n)^{1/2}$ , for some constant parameter,  $\delta \geq 0$ , with

$$\hat{\theta}_{ij} := \frac{1}{n} \sum_{l=1}^n ((X_{li} - \bar{X}_i)(X_{lj} - \bar{X}_j) - \sigma_{ij}^*)^2,$$

and where  $\bar{X}_i := \sum_{l=1}^n X_{il}/n$ .

For finite samples, the estimator  $\tilde{\Sigma}$  may not necessarily be a positive definite matrix. In this paper, we therefore use an algorithm due to Higham (2002) in order to locate a close positive definite matrix (see also Cheng and Higham, 1998). The resulting matrix, say  $\tilde{\Sigma}_{PD}$ , is then used in place of  $\tilde{\Sigma}$  in the test  $T_1$  above, and in place of the corresponding  $\hat{\Sigma}_j$  in the tests  $T_k$  above.

## 5. Simulation Studies

In this empirical study, we evaluate the statistical power of the two-sample test  $T_2$  for Laplacians, under different choices of number of vertices and for increasing sample sizes. We simulate network-based data for  $n$  subjects in each group, and focus our attention on two-sample experimental designs. Motivated by the neuroimaging application underlying the methodological development just described, the data generating process relies on (i) the selection of a network topology and the construction of an associated covariance matrix, (ii) the generation of multivariate time series for each network model, and (iii) the construction of subject-specific Laplacians based on either the covariance or the mutual information matrices.

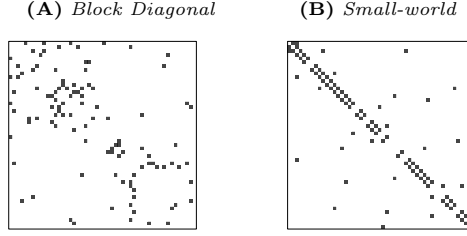
### 5.1. Network Topologies

In these simulations, we consider two types of network topology, specified through a binary matrix,  $A_1$  of order  $d \times d$ . Once the topology of the first sample is established, a second matrix,  $A_2$ , is constructed for the second sample, by randomly rewiring the original adjacency matrix. Firstly, we consider a block-diagonal structure for  $A_1$ , which represents the grouping of several vertices into two homogeneous communities, such that

$$A_1 := \begin{pmatrix} X & R \\ R & Y \end{pmatrix},$$

where  $X$  and  $Y$  are square matrices of dimensions  $\lceil d/2 \rceil$  and  $\lfloor d/2 \rfloor$ , respectively. The elements of  $X$  and  $Y$  are given a value of 1 according to independent Bernoulli variates with proportion  $p_1 := 4/d$ ; whereas the elements of  $R$  take a value of 1 with a probability of  $p_2 := 1/(2d)$ . These choices of  $p_1$  and  $p_2$  ensure that the corresponding block models are *sparse* in the sense that their numbers of edges are proportional to their numbers of vertices, as  $d$  grows.

Secondly, we specify a small-world network structure, by constructing a regular network with a ring topology, whose number of edges is taken to be proportional to  $d$ , which again enforces sparsity. The edges of this network are then randomly rewired (Watts and Strogatz, 1998). The choice of  $N_e$  is here motivated by a desire to maintain some level of comparison between the block-diagonal model and the small-world topology. Using such  $N_e$ 's, we ensure that both types of networks have approximately the same number of edges. These two



**Fig 2.** Simulated matrices over  $d = 50$  vertices. In (A) and (B), matrices with a block-diagonal structure and a small-world topology are respectively represented.

families of network topologies are illustrated in Figure 2 for simulated networks of size  $d = 50$ .

In both of these models, the group-specific covariance matrices,  $\Sigma_g$ 's, were then constructed using a mixture model, based on the binary matrices,  $A_g$ 's; with  $g = 1, 2$  denoting the group label for each independent sample. The diagonal elements of the  $\Sigma_g$ 's are given by

$$\Sigma_{aa,g} \stackrel{\text{iid}}{\sim} \exp(\lambda), \quad a = 1, \dots, d;$$

whereas the off-diagonal elements of the  $\Sigma_g$ 's are constrained by their corresponding adjacency matrices,  $A_g$ 's, as follows,

$$\Sigma_{ab,g} | A_{ab,g} \stackrel{\text{iid}}{\sim} |A_{ab,g} N(\mu_1, \sigma^2) + (1 - A_{ab,g}) N(\mu_2, \sigma^2)|;$$

for every  $a \neq b$ , and where the parameters of the mixture model are given the following values,  $\lambda := 4$ ,  $\mu_1 = 1$ ,  $\mu_2 = 0$  and  $\sigma^2 = .2$  for all simulation scenarios; thereby producing a high signal-to-noise ratio, permitting to distinguish between the different types of entries in the matrices  $\Sigma_g$ . Note that none of these simulation scenarios for  $\Sigma_g$  guarantees that the resulting  $\Sigma_g$  is positive definite. Consequently, we projected the resulting matrix to a close positive definite matrix, using the method described in Section 4.3). Once the  $\Sigma_g$ 's were obtained, they were fixed for each scenario, and used to generate different multivariate time series.

## 5.2. Noise Models

Resting-state or default-mode brain networks have been investigated by a large number of researchers in neuroimaging (Thirion et al., 2006; Beckmann et al., 2005). The main difficulty in simulating such networks stems from the absence of a prior to produce such resting-state patterns of activities (Leon et al., 2013; Kang et al., 2012). For each subject, we construct a set of  $d$  sequences of  $T$  realizations, where  $d$  represents the number of ROIs, and  $T$  denotes the total number of time points. These sequences are drawn from two different generating

processes. In the first scenario, these sequences of realizations are drawn from a multivariate Gaussian, such that the random vectors  $X_{itg} \in \mathbb{R}^d$  are given by

$$X_{itg} \stackrel{\text{iid}}{\sim} N_d(0, \Sigma_g), \quad \forall i = 1, \dots, n; t = 1, \dots, T; \quad (3)$$

where  $g = 1, 2$  denotes group affiliation. By contrast, in a second scenario, we model these sequences as multivariate time series, using an autoregressive process, of the form,

$$X_{itg} = \alpha + \varphi X_{i,t-1,g} + \epsilon_{itg}, \quad (4)$$

for every  $t = 1, \dots, T$ ; where  $\varphi \in \mathbb{R}$ , and  $\alpha, \epsilon_{tsg} \in \mathbb{R}^d$ . The first vector of this process is given by  $X_{i0g} = \alpha + \epsilon_{i0g}$ . For expediency, the autoregressive coefficient is set to be identical for all ROIs. Moreover, we restrict ourselves to autoregressive processes that are *wide-sense stationary*, by setting  $|\varphi| < 1$ . In this autoregressive model, the error terms are sampled from the  $d$ -dimensional normal distribution,  $\epsilon_{itg} \stackrel{\text{iid}}{\sim} N_d(0, \Sigma_g)$ , for every  $i = 1, \dots, n$  and  $t = 0, \dots, T$ . The analysis using the autoregressive model will be provided in the supplementary material.

### 5.3. Sample Estimators

For each synthetic data set, the subject-specific association matrices are computed. From these matrices we define the (weighted) network Laplacian matrices that form the ‘sample’ of interest. We consider either the covariance or the mutual information as an association measure. Both measures have been used in neuroimaging for constructing networks. But while the first yields adjacency matrices that are guaranteed to be positive semi-definite, the second does not (Jakobsen, 2014). Our framework accomodates both choices with equal ease.

When using covariances, we compute the subject-specific matrices,

$$S_{ig} := \frac{1}{T-1} \sum_{t=1}^T (X_{itg} - \hat{X}_{ig})(X_{itg} - \hat{X}_{ig})',$$

with  $\hat{X}_{ig} := T^{-1} \sum_{t=1}^T X_{itg}$ . Alternately, for the mutual information, we have for each subject a matrix,  $S_{ig}$ , whose entries take the form,  $s_{ig,ab} := I(X_{iag}, X_{ibg})$ , for every  $1 \leq a, b \leq d$ , where the mutual information is defined for every pair of discrete random variables  $X$  and  $Y$  with respective codomain  $\mathcal{X}$  and  $\mathcal{Y}$  as follows,

$$I(X, Y) := \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right),$$

Note that the mutual information is here computed for continuous random variables. Thus, we are using a discretization of the original range of the time series, as described by Dougherty, Kohavi and Sahami (1995).



The weighted combinatorial Laplacian of each sample association matrix is then given for every  $i^{\text{th}}$  subject in the  $g^{\text{th}}$  experimental group by,

$$L_{ig} := D(S_{ig}) - S_{ig},$$

where  $D(S_{ig})$  is a diagonal matrix of weighted degrees, with non-zero entries given by  $\{D(S_{ig})\}_{aa} := \sum_{b=1}^d s_{ig,ab}$ , for every  $a = 1, \dots, d$ .

The target parameters of interest are here the combinatorial Laplacians of the unknown covariance matrices,  $L_g := D(\Sigma_g) - \Sigma_g$ . This unknown quantity is estimated using the following sample mean Laplacian,

$$\hat{L}_g := \frac{1}{n} \sum_{i=1}^n L_{ig},$$

for each group. This estimator is linearly related to the sample mean of the sample covariance matrices,  $\hat{S}_g := n^{-1} \sum_{i=1}^n S_{ig}$ , by the following relation,  $\hat{L}_g = D(\hat{S}_g) - \hat{S}_g$ . The second moments of the group-specific combinatorial Laplacians are the following sample covariance matrices,

$$\hat{\Xi}_g := \frac{1}{n-1} \sum_{i=1}^n (\phi(L_{ig}) - \phi(\hat{L}_g))(\phi(L_{ig}) - \phi(\hat{L}_g))',$$

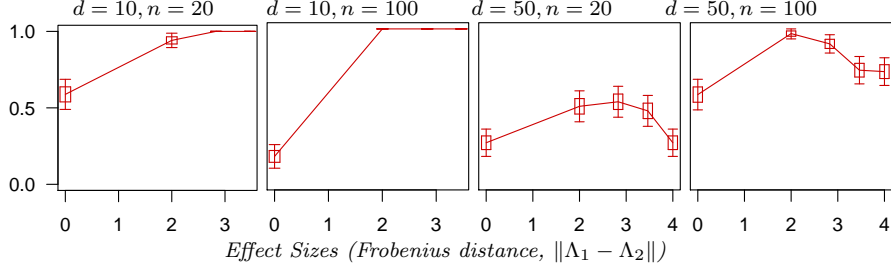
for  $g = 1, 2$ . These sample covariance moments are then modified using the covariance estimation techniques described in Section 4.3.

#### 5.4. Simulation Design

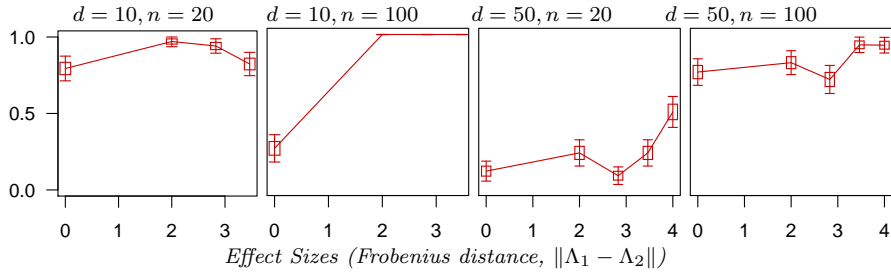
Four main factors were made to vary in this set of simulations. In line with the subsequent real-data analysis, we considered sample sizes of  $n = 20, 100$  per group. This was deemed representative of the number of subjects found in most neuroimaging studies. Secondly, we varied the network sizes, with  $d$  taking values of 10 and 50, corresponding to what would result from coarser and finer definitions of regions of interest (ROIs) in practice. This range of network sizes allowed us to identify the effect of network size on the statistical power of our test. Larger dimensions were expected to decrease power.

In each of these scenarios, we computed the statistical power of the two-sample tests, using different effect sizes. Here, the effect size was defined as the Frobenius distance between the two population means. The effect size of the test was varied by rewiring the population means, thereby increasing the differences between the two groups. These repeated rewiring resulted in differences between the population means,  $\Lambda_1$  and  $\Lambda_2$ , which will be represented by the Frobenius distance,  $\|\Lambda_1 - \Lambda_2\|_F$ , in the ensuing discussion. For each set of conditions, the simulations were repeated 100 times in order to obtain an empirical estimate of the theoretical power of the two-sample test statistic for Laplacians, under these conditions.

**(A) Block Model**



**(B) Small-world Model**



**Fig 3.** Power curves for the simulated two-sample tests using the *covariance* estimation procedure, under a multivariate Gaussian model, with error bars based on one and two standard errors from the mean. The  $y$ -axis indicates the probability of rejecting the null hypothesis when it is false; whereas the  $x$ -axis is a proxy measure of effect size, computed using the Frobenius distance between the two population means,  $\|\Lambda_1 - \Lambda_2\|$ . These results are presented for networks on  $d = 10$  and  $50$  vertices, with group sizes of  $n = 20, 100$ , and over  $T = 200$  time points, and based on 100 iterations per condition with respect to the block and small-world topologies.

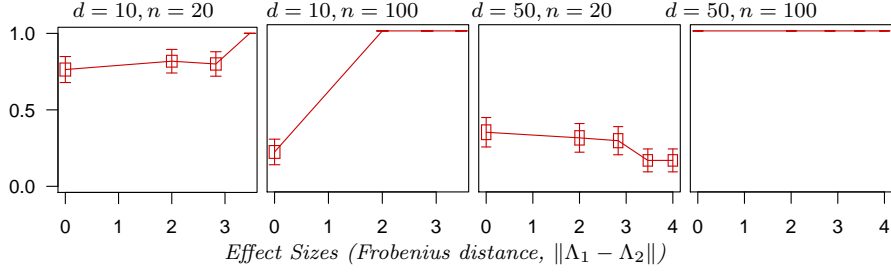
### 5.5. Simulation Results

The results of these simulations are reported in Figures 3 and 4 for the choices of covariance and mutual information procedures, respectively, in defining networks from the underlying basic measurements. Larger power for a given effect size is better. Observe that the power curves are ‘roughly’ increasing with power size<sup>1</sup>. These results correspond to the Gaussian noise model. Comparable results for the AR noise model can be found in the supplementary material.

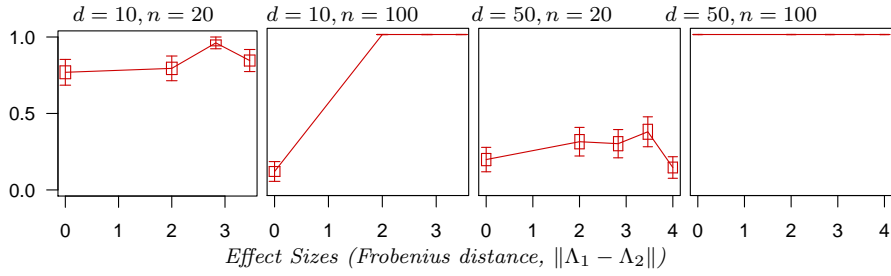
When considering networks defined through covariances, the power of the two-sample test for Laplacians was found to be empirically well-behaved, when  $d = 10$  and  $n = 100$ . This was true for both the block and small-world topological models, as illustrated in the second column of plots, in Figure 3. The statistical power, however, was poor under the two topological models, for small sample

<sup>1</sup>We are here solely using a *proxy* measure of the effect sizes (i.e.  $\|\Lambda_1 - \Lambda_2\|$ ). Since the true covariance matrix is unknown, such proxy measures are therefore not normalized.

**(A) Block Model**



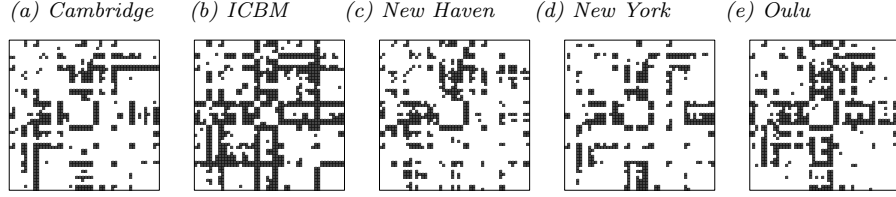
**(B) Small-world Model**



**Fig 4.** Power curves for the simulated two-sample tests using the *mutual information* estimation procedure, under a multivariate Gaussian model, with error bars based on one and two standard errors from the mean. The simulation parameters are identical to the ones described in Figure 3.

sizes. With  $n = 20$  subjects in each group, the test performed poorly both in terms of rejecting the null hypothesis, when it was incorrect and in terms of accepting it when it was true. When increasing the size of the networks of interest to  $d = 50$ , the probability of rejecting  $H_0$ , when it was false remained satisfactorily high. However, increasing the size of the networks of interest resulted in a higher likelihood of committing a type I error (i.e. rejecting  $H_0$ , when  $H_0$  is true), as can be seen from the last column of plots in Figures 3(a) and 3(b).

The use of the mutual information in defining networks provided better results for small network sizes. While the behavior of the two-sample test for small sample sizes, i.e.  $n = 20$ , remained poor, it greatly benefited from an increase in sample size. Albeit using networks defined through mutual information seemed to exhibit slightly less statistical power, under the alternative hypothesis, it resulted in lower type I error. However, when considering large networks with  $d = 50$ , the mutual information failed to distinguish between scenarios under  $H_0$  and scenarios under the alternative hypothesis. Thus, while the mutual information may be recommended in practice for small network sizes, our results here suggest that covariance estimation should generally be preferred for larger networks.



**Fig 5.** Mean Laplacians for five subsamples of functional neuroimaging networks in the 1000 FCP data set, corresponding to five collecting sites, including data from (a) Harvard University, (b) the International Consortium for Brain Imaging (ICBM), (c) Yale University, (d) New York University, and (e) the University of Oulu in Finland. These five groups respectively contained 198, 257, 63, 59, and 103 subjects, respectively. The Laplacians have been binarized as in Figure 1.

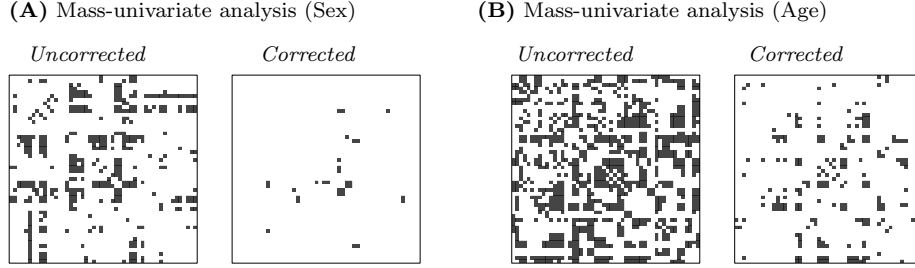
## 6. Analysis of the 1000 FCP Data Set

Different aspects of the 1000 FCP data set were considered. Firstly, we use a one-sample test for comparing the Laplacian mean to a subsample of the data. We then tested for sex and age differences using the two- and  $k$ -sample tests for Laplacians. Finally, we analyzed the differences between connectivity patterns of five subgroups of subjects from five different collection centers. After excluding subjects for which demographics data were incomplete, we analyzed  $n = 1017$  subjects.

### 6.1. Inference on Full Data Set

As described in Section 1.2, the 1000 FCP data provides a unique opportunity for neuroscientists to extract a reference template of human connectivity. We tested the reliability of that template using a one-sample Laplacian test for some random subsample of the data. We computed the reference mean Laplacian over the full FCP sample, which is here treated as a *population parameter*,  $\Lambda_0$ . This was compared with a given subsample of size  $n = 100$ . We tested for the null hypothesis that the sample mean,  $\bar{L}_1$ , was equal to the reference mean  $\Lambda_0$ . This hypothesis was rejected with high probability ( $T_1 > 10^4$ ).

The partitioning of the 1000 FCP data set by sex is provided in Figure 1(A). As highlighted in the introduction, sex differences have been found to influence patterns of brain connectivity at a level that is discernible using neuroimaging data. Here, we tested whether such sex differences were significant using the two-sample test for Laplacians. The null hypothesis of no group differences was rejected with high probability ( $T_2 > 10^6$ ). Subjects in the 1000 FCP database can also be grouped according to age. In Figure 1(B), we have divided the FCP sample into three subgroups of approximately equal sizes, with 386, 297, and 334 subjects; for subjects younger than 22, between 22 and 32, and older than



**Fig 6.** Mass-univariate analyses were conducted to test for local differences in connectivity due to sex and age in the full FCP data set. In each case,  $\binom{d}{2}$  tests were performed independently for each of the off-diagonal entries in the Laplacians. Sex differences and age differences are reported in panel (A) and (B), respectively. In each case, the first matrix denotes the entries that were found to be significantly different between the groups at  $\alpha = .05$ ; whereas the second matrix represents the significant entries after Bonferroni correction. Black denotes significant entries.

32, respectively. The  $k$ -sample Laplacian test was performed to evaluate the hypothesis stating that these  $k = 3$  groups were drawn from the same population. This null hypothesis was also rejected with high probability ( $T_3 > 10^6$ ). These results should be compared with the use of a mass-univariate approach, in which a single hypothesis test is run for each voxel. The significant voxel-level differences detected using a mass-univariate approach for sex and age, are reported in Figure 6.

The 1000 FCP is based on a consortium of universities from around the world. Despite the best efforts to coordinate and standardize the data collection process, some differences may still exist between the mean connectivity patterns of each site-specific sample. It is therefore natural to test whether these subsamples were drawn from the same population. We here focused on the five largest collection sites, which included Harvard University, the International Consortium for Brain Imaging (ICBM), Yale University, New York University and the University of Oulu in Finland. The mean Laplacians for each of these sub-samples are reported in Figure 5. These five groups respectively contained 198, 257, 63, 59, and 103 subjects. Using the  $k$ -mean test described in Section 4.2, we found that the null hypothesis stating that all such site-specific means were identical—that is,  $H_0 : \Lambda_1 = \dots = \Lambda_5$ , was rejected with high probability ( $T_5 > 10^{10}$ ).

## 6.2. Inference on Partial Data Set

The results of the previous section were compared with another analysis based on a small subset of connectomes. The 1000 FCP data set is indeed exceptionally large for the field of neuroimaging. By contrast, most papers using MRI data tend to report results based on smaller data sets, usually containing in the region of 20 subjects. Here, we have replicated the various statistical tests described in the last section for such a small sample size, in order to produce an analysis

more reflective of what might be performed by, say, a single lab. This small subset of subjects were selected randomly, but the findings were found to be consistent for different choices of random sub-samples.

The conclusions of the network-level tests for the different hypotheses of interest were found to be robust to a large decrease in sample size. As for the larger data set, sex differences were also found to be highly significant ( $T_2 > 10^9$ ), when solely considering 10 female and 10 male subjects. Similarly, for the one-sample test evaluating whether the mean Laplacian of interest was significantly different from a reference Laplacian (i.e. the mean Laplacian in the full 1000 FCP data set), was found to be very significant ( $T_1 > 10^{10}$ ). When comparing the three different age cohorts with 10 subjects in each category, we also rejected the null hypothesis with high probability ( $T_3 > 10^{10}$ ). Finally, for several sites, a re-analysis based on 10 subjects per site showed that the mean Laplacians extracted from the different sites was highly likely to have been drawn from different populations ( $T_3 > 10^{10}$ ).

These significant results should be contrasted with the use of a mass-univariate approach, in this context. We compared the conclusions of a network-level Laplacian test for sex, with the ones of a mass-univariate approach based on 10 female and 10 male subjects. No local differences were here found, even prior to correct for multiple comparisons. This highlights one of the important advantages of using a global test in this context. While the mass-univariate approach fails to detect any sex differences at the local level, our proposed global test, by contrast, has sufficient power to reject the null hypothesis at the global level.

## 7. Discussion

In this paper, we have analyzed a large neuroimaging data set, using a novel framework for network-based statistical testing. The development of this framework is grounded in a formal asymptotic theory for network averages, developed within the context of a well-defined notion of the space of all Laplacians corresponding to the networks. Importantly, we have showed that using the global tests that result from our framework may provide the researcher with decidedly more statistical power than when using a mass-univariate approach, which is the standard approach in the field.

To the best of our knowledge, we are the first to ascribe a notion of a ‘space’ to the collection of graph Laplacians and to describe the geometrical properties of this space. While we have found it convenient for the purposes of exposition simply to summarize these results in the main body of the paper, and to collect details in the appendices, it is important to note that this initial step is crucial in allowing us to bring to bear recent probabilistic developments in the field of shape analysis to produce our key central limit theorem, upon which the distribution theory for our tests lies. We note too that the framework we offer is quite general and should, therefore, as a result be quite broadly applicable. Nevertheless, this initial work also has various limitations, and furthermore sets the stage for numerous directions for extensions, which we describe briefly below.

### 7.1. Limitations

It can be expected that there be a tradeoff in the performance of our tests between sample size  $n$  and the dimension  $d$  of the networks in the sample. This expectation is confirmed in our simulations, where one can observe that for a given sample size  $n$ , the rate of type I error increases beyond the nominal rate, as  $d$  increases. Since our test can be seen to be equivalent to a Hotelling  $T^2$  on the off-diagonal elements of the Laplacians, it follows that sample sizes of order  $O(d^2)$  would be required to control for this increase in type I error rate. For the analysis of the full FCP data set, this condition was approximately satisfied, since this data set contains more than 1000 subjects, and we were here comparing networks with 50 vertices. In their current forms, such global statistical tests may therefore be most applicable to very large data sets, or to relatively small networks. However, our analysis of the smaller subsets of the FCP data (i.e., mimicking analysis at the level of a single lab) suggests that even at low sample sizes the test is well-powered against the alternative of differences in network group averages.

Computationally, the method employed in this paper was also challenging since the application of the Laplacian test required the inversion of a large covariance matrix. We have here resorted to different methods to facilitate this process including the use of modern sparse estimation techniques (Cai and Liu, 2011), as well as the modification of the resulting sample covariance matrix estimates in order to force positive definiteness (Cheng and Higham, 1998; Higham, 2002). Practically, however, such methods remain computationally expensive, and may therefore limit the size of the networks that one may wish to consider when using such Laplacian tests.

### 7.2. Extensions

In our work here (specifically, as described in Section 3) we show that the ‘space’ of networks – *without any structural constraints* – behaves ‘nicely’ from the mathematical perspective, and therefore we are able to develop a corresponding probability theory and statistical methods for one- and two-sample assessment of network data objects. However, one of the most fundamental results that has emerged from the past 20 years of complex network research is the understanding that real-world networks typically (although not exclusively) in fact tend to possess a handful of quite marked structural characteristics. For example, most networks are relatively sparse, in the sense that the number of edges is on the same order of magnitude as the number of vertices. Other common key properties include heterogeneous degree distributions, cohesive subgraphs (a.k.a. communities), and small-world behavior (see Newman, 2010, chap.8).

The ubiquity of such characteristics in real-world networks has been well-established. Importantly, this fact suggests that the appropriate (differential or metric measure) geometry of the ‘space of all networks’ – or, more formally, the space of Laplacians corresponding to such networks – depends on the constraints

imposed on these networks/Laplacians. In particular, other choices of network constraints can lead to metric geometry problems embedded inside Riemannian geometry problems. For example, imposing sparseness on a network leads to nontrivial geometry. The Euclidean average of two sparse networks/matrices need not be sparse, and apart from simple scalings, one expects the set  $\mathcal{L}$  of sparse matrices, properly defined, to be a discrete subset of the manifold of positive semi-definite matrices (PSD) and hence far from convex. Thus it is natural to define the average of two sparse matrices to be the sparse matrix closest to the Euclidean average, but this may be computationally unappealing. Moreover, the Riemannian measure on PSD does not determine a measure on  $\mathcal{L}$ , so computing Fréchet means becomes problematic. Of course, one can impose a uniform distribution on  $\mathcal{L}$ , but this risks losing all geometric relations between  $\mathcal{L}$  and PSD. Hence, there are a variety of open problems to be studied examining the implications of network structural constraints on the space  $\mathcal{L}$ .

Furthermore, since the asymptotic theory we exploit from shape analysis relies heavily on the topological and geometrical properties of the space within which they are brought to bear, we can expect that different network constraints will require different levels of effort in producing central limit theorems. More precisely, while a general asymptotic distribution theory for Fréchet means in metric spaces has recently been derived by Bhattacharya and Lin (2013), this theory requires that a number of conditions be satisfied, the verification of which can be expected to become increasingly difficult as the geometry of the space becomes complicated. Thus, accompanying the various extensions in geometry described above are likely to be corresponding challenges in probability theory and shape analysis.

Finally, while the 1000 FCP data set is unique in its magnitude and richness, which in turn has allowed us to pose and answer a good number of questions relevant to neuroscience in the analyses using our proposed testing framework, there remains much additional empirical work to be done applying our methods, in order to more fully establish both their capabilities and their limitations. We would anticipate that with the recently started BRAIN initiative, and other endeavors like it, that within five years there will be a plethora of databases of network-based objects in neuroscience, providing more than ample motivation not only for the further testing of methods like the ones we have proposed here, but also for extending other tools from classical statistics to network data.

## Appendix A: Proofs from Section 3.1

PROOF OF THEOREM 1. Let the matrix  $E$  of order  $d \times d$  be partitioned in the following manner,

$$E = \begin{pmatrix} d-1 & 1 \\ A & v \\ v' & x \end{pmatrix} \quad \begin{matrix} d-1 \\ 1 \end{matrix}$$

This matrix is assumed to satisfy conditions (1), (2), and (4). We will call the set of such matrices  $\mathcal{T}$ . Assume that  $A$ , the top left  $(d-1) \times (d-1)$  block of  $E$ , has nonzero determinant. We want to show that some  $d(d-1)/2$ -dimensional ball around



$E$  continues to lie in  $\mathcal{T}$ . Since the rank of  $E$  is  $d - 1$ , the last column of  $E$  is a linear combination of the first  $d - 1$  columns. Since the columns of  $E$  add to zero and  $E$  is symmetric, the rows of  $E$  add to zero. For  $v' = (v_1, \dots, v_{d-1})$ , we must have

$$v_i = -\sum_{j=1}^{d-1} A_{ij}, \quad \text{and} \quad x = -\sum_{j=1}^{d-1} v_j.$$

Thus,  $v$  and  $x$  are determined by the entries of  $A$ .

The matrix,  $A$ , is symmetric. Thus, it lies in the subspace  $S$  of  $\mathbb{R}^{(d-1)^2}$  of dimension  $d(d-1)/2$  consisting of symmetric matrices.  $\text{Det}(A) \neq 0$ , so for some matrix  $A_\epsilon$  in some small neighborhood  $U$  of  $A$  in  $S$ ,  $\text{det}(A_\epsilon) \neq 0$ . Each choice of  $A_\epsilon$  determines a corresponding  $v$  and  $x$ . Conversely, each  $E_\epsilon \in \mathcal{T}$  sufficiently close to  $E$  in the  $\mathbb{R}^{d^2}$  norm has  $\text{det}(A_\epsilon) \neq 0$  and  $A_\epsilon - A = X$  is symmetric, so  $A_\epsilon$  and hence  $E_\epsilon$  is determined by  $X$ . Thus, a neighborhood of  $E$  in  $\mathcal{T}$  is bijective to  $U$ . It is easy to check that this bijection is a diffeomorphism.

If some other  $(d-1) \times (d-1)$  block  $B$  of  $E$  has nonzero determinant, we note that the top  $(d-1) \times (d-1)$  block  $A$  of the matrix determines the entire matrix as above. Any small symmetric perturbation  $A_\epsilon$  of  $A$  (with the necessary perturbations of the last row and column to preserve (4)) still satisfies  $\text{det}(B_\epsilon) \neq 0$ . Conversely, any  $B \in \mathcal{T}$  sufficiently close to  $E$  so that  $\text{det}(B_\epsilon) \neq 0$  determines a symmetric perturbation of  $A$  as above. Hence, we again obtain a neighborhood of  $E$  in  $\mathcal{T}$  parametrized by a neighborhood  $U$  of  $A$  in  $S$ . This shows that  $\mathcal{T}$  is a submanifold of  $\mathbb{R}^{d^2}$  of dimension  $d(d-1)/2$ . The set of matrices satisfying (5) alone is an open convex cone in  $\mathbb{R}^{d^2}$ . When we intersect the submanifold  $\mathcal{T}$  with this cone, we get an open submanifold  $\mathcal{T}'$  of  $\mathcal{T}$ . Thus  $\mathcal{T}'$ , the set of matrices with (1), (2), (4), (5), is also a submanifold of  $\mathbb{R}^{d^2}$  of dimension  $d(d-1)/2$ .

The space  $\mathcal{T}'$  has several connected components. A matrix  $E_0$  with  $k$  positive eigenvalues and a matrix  $E_1$  with  $k' \neq k$  positive eigenvalues lie in different components, as a path in  $\mathcal{T}'$  from  $E_0$  to  $E_1$  would contain a matrix with a zero eigenspace of multiplicity at least two. Conversely, if  $k = k'$ , then  $E_0$  and  $E_1$  are in the same component of  $\mathcal{T}'$ . For the line segment  $E_t = (1-t)E_0 + tE_1$  stays in  $\mathcal{T}'$ , for every  $t \in [0, 1]$ . Since the components are open, the component of  $\mathcal{T}'$  satisfying  $k = d-1$  is again a submanifold of dimension  $d(d-1)/2$ . But this component has condition (3), and so is precisely  $\mathcal{L}_d$ . This proves that  $\mathcal{L}_d$  is a manifold of dimension  $d(d-1)/2$ .

For the convexity statement, conditions (2) – (5) are convex conditions; e.g. for (3), if  $A$  and  $B$  are positive semidefinite, then

$$\langle (tA + (1-t)B)v, v \rangle = t\langle Av, v \rangle + (1-t)\langle Bv, v \rangle \geq 0$$

for  $t \in [0, 1]$  and  $v \neq 0$ . Clearly, (1) – (5) together is a convex condition. For if  $A$  and  $B$  satisfy (1) – (5), then  $A$  and  $B$  come from weighted connected graphs, as does  $tA + (1-t)B$ . Since a graph is connected iff the rank of the corresponding Laplacian matrix has rank  $d-1$ , the rank of  $tA + (1-t)B$  is  $d-1$  for  $t \in [0, 1]$ . Thus  $\mathcal{L}_d$  is a convex submanifold of  $\mathbb{R}^{d^2}$ .

To show that  $\mathcal{L}_d$  lies in an affine subset, fix  $E \in \mathcal{L}_d$ . For  $k = d(d-1)/2$ , take  $k$  distinct points  $s_i$  in  $\mathcal{L}_d$ , none of them equal to  $E$ , such that the convex hull of these points contains  $E$ . (For example, two of the points can be close to  $E \pm S$  for a small symmetric matrix  $S$ .) For generic choices, the  $k$  points plus  $E$  determine an (affine)  $k$ -plane  $P$ , and the convex hull of these points lies in both  $P$  and  $\mathcal{S}$ . Since  $P$  and  $\mathcal{L}_d$  have the same dimension, the open convex hull is exactly a neighborhood of  $E$  in  $\mathcal{L}_d$ .

We now show that the plane  $P$  is independent of the choice of  $E$ . Since  $\mathcal{L}_d$  is convex, it is connected. Take  $F \in \mathcal{S}$ , let  $\ell$  be the Euclidean line segment from  $E$  to  $F$ , and set  $E_t = (1-t)E + tF \in \mathcal{L}_d$ . Arguing as above, we find a plane  $P_t$  containing a neighborhood  $V_t$  of  $E_t$  in  $\mathcal{L}_d$ . By compactness, there exist  $0 = t_0, \dots, t_n = 1$  with  $\cup_{i=0}^n V_{t_i} \supset \ell$ . If  $P = P_0 \neq P_{t_1}$ , then some line segment from one of the  $s_i$ 's determining  $P_0$  to one of the  $s_j$ 's determining  $P_{t_1}$  does not lie in  $\mathcal{L}_d$ , a contradiction. Thus  $P = P_{t_1}$ , and by induction,  $P = P_1$ . Since  $F$  is arbitrary in  $\mathcal{L}_d$ , it follows that  $\mathcal{L}_d$  lies in  $P$ .  $\square$

PROOF OF COROLLARY 1. In the notation of the proof of Theorem 1, assume that  $E$  has conditions (1), (2), (4), (5'). Then  $A$  is symmetric and has  $a_{ij} \leq 0$ . Thus  $A$  is in bijection with the closed "quadrant"  $\{(x^1, \dots, x^{d(d-1)/2}) : x^i \leq 0\}$ , which is the basic example of a manifold with corners. If the rank  $d-1$  submatrix  $B$  of  $E$  is not in the top left corner, a relabeling of coordinates moves  $B$  to the top left corner. Since the relabeling takes the closed quadrant to a closed quadrant, a neighborhood of  $B$  has the structure of a manifold with corners. It is trivial to check that transition maps from chart to chart are smooth. If we impose (3), then as in the previous proof we pick out one connected component of this manifold with corners, and each component is a manifold with corners. The statements on convexity and affine subspaces follow immediately from Theorem 1, since  $\mathcal{L}'_d$  is a dense subset of  $\mathcal{L}_d$ .  $\square$

PROOF OF THEOREM 2. Assume the  $\ell \times \ell$  block with nonzero determinant occurs in the top left corner; the other cases are handled as in the proof of Theorem 1. Thus let

$$E = \begin{array}{c} \ell \qquad d-\ell \\ \left( \begin{array}{c|ccc} A & v_1 & \dots & v_{d-\ell} \\ \hline v'_1 & & & \\ \vdots & & & \\ v'_{d-\ell} & b_1 & \dots & b_{d-\ell} \end{array} \right) \begin{array}{c} \ell \\ \\ \\ d-\ell \end{array} \end{array}$$

have conditions (1 $_\ell$ ), (2), (4). Here,  $v_i$  is an  $\ell \times 1$  column vector, and  $b_i$  is a  $(d-\ell) \times 1$  column vector. The dimension of the set of  $\ell \times \ell$  symmetric matrices  $A$  with nonzero determinant is  $\ell(\ell+1)/2$ . Since the last  $d-\ell$  columns must be linear combinations of the first  $\ell$  columns, we have

$$v_i = \sum_{j=1}^{\ell} v_{ij} a_j, \quad i \in \{1, \dots, d-\ell\};$$

where  $a_j$  is the  $j^{\text{th}}$  column of  $A$ . The  $v_{ij}$ 's are arbitrary for  $i = 1, \dots, d-\ell-1$ , but (4) implies that the  $v_{d-\ell,j}$ 's are determined by the previous  $v_{ij}$ 's. Therefore, we get another  $(d-\ell-1)\ell$  degrees of freedom (i.e. dimensions), so the dimension of the space of matrices with (1 $_\ell$ ), (2), (4) is  $\ell(\ell+1)/2 + (d-\ell-1)\ell = d\ell - \ell(\ell+1)/2$ . The argument for adding in conditions (3) and (5) goes as before.  $\square$

### Appendix B: Proof of Theorem 3

The Laplacian CLT considered in this paper is a specialization of a general result due to Bhattacharya and Lin (2013), which considers a metric space  $(\mathcal{X}, \rho)$  equipped with

a probability measure  $Q$ . In addition to the conditions stated in the main body of the paper, two further regularity assumptions must be made on the first and second derivatives of the function  $\rho^2(\phi^{-1}(u), x)$ . These conditions are described below as (A5) and (A6).

Bhattacharya and Lin (2013) have shown that Euclidean coordinates of a Fréchet mean defined on a metric space converges to a normal distribution, under the following assumptions: (A1) the Fréchet mean  $\mu$ , as described in equation (1) is unique; (A2)  $\mu \in A \subseteq \mathcal{X}$ , where  $A$  is  $Q$ -measurable, and  $\hat{\mu}_n \in A$ , almost surely; (A3) there exists a homeomorphism  $\phi : A \rightarrow U$ , for some  $s \geq 1$ , where  $U$  is an open subset of  $\mathbb{R}^s$ ; (A4) for every  $u \in U$ , the map,  $u \mapsto h(u; x) := \rho^2(\phi^{-1}(u), x)$ , is twice differentiable on  $U$ , for every  $x \in \mathcal{X}$  outside a  $Q$ -null set; (A5) for every pair  $1 \leq k, l \leq s$ , with  $u \in U \subseteq \mathbb{R}^s$  and  $x \in \mathcal{X}$ , letting

$$D_k h(u; x) := \frac{\partial}{\partial u_k} h(u; x), \quad \text{and} \quad D_{k,l} h(u; x) := \frac{\partial^2}{\partial u_k \partial u_l} h(u; x),$$

we require that  $\mathbb{E}[|D_k h(u; x)|^2] < \infty$ , and  $\mathbb{E}[|D_{k,l} h(u; x)|] < \infty$ ; moreover, (A6) defining  $f_{k,l}(\epsilon, x) := \sup\{|D_{k,l} h(u; x) - D_{k,l} h(\phi(\mu); x)| : |u - \phi(\mu)| < \epsilon\}$ , we also require modulus continuity, such that  $\mathbb{E}[|f_{k,l}(\epsilon; Y)|] \rightarrow 0$ , as  $\epsilon \rightarrow 0$ , for every  $1 \leq k, l \leq s$ ; and finally, (A7) the matrix,  $B := \{\mathbb{E}[D_{k,l} h(\phi(\mu); Y)]\}_{k,l=1,\dots,s}$ , should be non-singular. Under these conditions, it is then true that the following convergence in distribution holds,

$$n^{1/2} (\phi(\hat{\mu}_n) - \phi(\mu)) \rightarrow N(0, B^{-1} V B^{-T}),$$

where  $V := \text{Cov}[D h(\phi(\mu); Y)]$  is assumed to be non-singular.

In our setting, we have drawn an iid sample of combinatorial Laplacians from an unknown generating distribution, such that we have  $Y_i \sim F(\Lambda, \Sigma)$ , for every  $i = 1, \dots, n$ , where  $\Lambda$  and  $\Sigma$  are the mean Laplacian and the covariance matrix of the upper triangle of  $Y$ , with respect to some unknown distribution,  $F$ . Observe that the space of interest is here  $\mathcal{L}'_d$ , equipped with the Frobenius distance, as stated in Corollary 1, thereby forming the metric space,  $(\mathcal{L}'_d, \|\cdot\|_F)$ . We will see that conditions (A1) – (A4) as well as (A7) are necessarily satisfied in our context. Moreover, we will assume that conditions (A5) and (A6) also hold.

Condition (A1) is readily satisfied, since we have demonstrated that the space of interest,  $\mathcal{L}'_d$ , is a convex subspace of  $\mathbb{R}^{d^2}$ ; and moreover the arithmetic mean is a convex function on that space by corollary 1. Thus, the sample Fréchet mean,  $\hat{L}_n$ , is unique, for every  $n \in \mathbb{N}$ . Secondly, we have assumed that the underlying measure gives a non-zero positive probability to a subset  $U \in \mathbb{R}^{d^2}$ , which contains  $\Lambda$ . Therefore, condition (A2) is satisfied, in the sense, that there exists a subset  $A \subseteq \mathbb{M}_{d,d}(\mathbb{R}^+)$ , such that  $A$  is  $\mathbb{P}$ -measurable. In addition, since the strong law of large numbers holds for the Fréchet mean (see Ziezold, 1977), we also know that  $\hat{L}_n \rightarrow \Lambda$ , almost surely; and therefore,  $\mathbb{P}[\hat{L}_n \in A] \rightarrow 1$ , as  $n \rightarrow \infty$ , as required by condition (A2).

For condition (A3), observe that, in our context, the homeomorphism of interest,  $\phi : A \mapsto U$ , is the *half-vectorization* function. This takes a matrix in  $\mathcal{L}'_d$ , and returns a vector in  $\mathbb{R}^{\binom{d}{2}}$ , such that for every  $Y \in \mathcal{L}'_d$ ,  $\phi(Y) := \text{vech}(Y)$ . Specifically, this vectorization is defined by a change of indices, such that for every  $i \leq j$ , with  $1 \leq i, j \leq d$ , we have  $[\phi(Y)]_{k(i,j)} := y_{ij}$ , with  $k(i,j) := (i-1)d + j$ . The inverse function,  $\phi^{-1}$ , is then readily obtained for every  $u \in U \subseteq \mathbb{R}^{\binom{d}{2}}$ , satisfying  $\phi^{-1}(u) = Y$ , as  $[\phi^{-1}(u)]_{ij} = y_{ij}$ . The bicontinuity of  $\phi$  is hence trivially verified and this map is therefore a homeomorphism.

For condition (A4), the function  $h(u; Y) := \rho^2(\phi^{-1}(u), Y)$ , for every  $u \in U \subseteq \mathbb{R}^{(d)}$  and every  $Y \in \mathcal{L}'_d$ , outside of a  $Q$ -null set, is here defined as

$$h(u; Y) := \|\phi^{-1}(u) - Y\|_F^2 = \sum_{i \leq j}^d \left( [\phi^{-1}(u)]_{ij} - y_{ij} \right)^2,$$

where the sum is taken over all the pairs of indices  $1 \leq i, j \leq d$ , satisfying  $i \leq j$ . The first derivative of this map with respect to the coordinates of the elements of  $\mathcal{L}'_d$  in  $\mathbb{R}^{(d)}$ , is straightforwardly obtained. Setting  $X := \phi^{-1}(u)$ , we have

$$D_{k(i,j)} h(u; Y) := \frac{\partial}{\partial u_{k(i,j)}} \|\phi^{-1}(u) - Y\|_F^2 = 2(x_{ij} - y_{ij}).$$

The second derivative of  $h(u; Y)$  can be similarly derived for every quadruple,  $1 \leq i, j, i', j' \leq d$ , satisfying  $k(i, j) \neq k(i', j')$ . When expressed with respect to  $\Lambda \in U$ , this gives

$$D_{k(i,j), k(i',j')} h(\phi(\Lambda); Y) = \begin{cases} 2, & \text{if } k(i, j) = k(i', j'), \\ 0, & \text{otherwise.} \end{cases}$$

It immediately follows that the matrix of second derivatives is  $B = 2I$ , and hence condition (A4) is verified. In addition, we have assumed that conditions (A5) and (A6) hold in our context. Finally, we have seen that the matrix  $B$  is diagonal and hence non-singular, as required by condition (A7).

We can also compute the covariance matrix of the resulting multivariate normal distribution. For this, we require the matrix  $V := \mathbb{Cov}[D h(\phi(\Lambda); Y)]$ . Given our choice of  $\phi$ , we need to consider the mean vector of  $D h(\phi(\Lambda); Y)$ , which is given for every  $1 \leq i, j \leq n$  by  $\mathbb{E}[D_{k(i,j)} h(\phi(\Lambda); Y)] = 2(\Lambda_{ij} - \mathbb{E}[Y]_{ij}) = 0$ . We can then compute the elements of  $V$ . For every quadruple  $1 \leq i, j, i', j' \leq n$ , this gives

$$\begin{aligned} V_{k(i,j), k(i',j')} &= \mathbb{E}[D_{k(i,j)} h(\phi(\Lambda); Y) \cdot D_{k(i',j')} h(\phi(\Lambda); Y)] \\ &= 4\mathbb{E}[(\Lambda_{ij} - Y_{ij})(\Lambda_{i'j'} - Y_{i'j'})] \\ &= 4(\mathbb{E}[Y_{ij}Y_{i'j'}] - \Lambda_{ij}\Lambda_{i'j'}), \end{aligned}$$

since the cross-term vanishes, after taking the expectation. Therefore, the asymptotic covariance matrix in Theorem 3 is indeed equal to the covariance matrix of the distribution, from which the  $Y_i$ 's have been sampled. That is, this covariance matrix is given by  $B^{-1}VB^{-T} = (2I)^{-1}V(2I)^{-1} = \mathbb{V}\text{ar}[\phi(Y)] = \Sigma$ . Therefore, all the conditions of Theorem 2.1 of Bhattacharya and Lin (2013) have been satisfied, and hence  $n^{1/2}(\phi(\hat{L}_n) - \phi(\Lambda)) \rightarrow N(0, \Sigma)$ , as stated in theorem 3.

## References

- ACHARD, S., SALVADOR, R., WHITCHER, B., SUCKLING, J. and BULLMORE, E. (2006). A Resilient, Low-Frequency, Small-World Human Brain Functional Network with Highly Connected Association Cortical Hubs. *J. Neurosci.* **26** 63–72.
- ARSIGNY, V., FILLARD, P., PENNEC, X. and AYACHE, N. (2007). Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications* **29** 328–347.

- AYDIN, B., PATAKI, G., WANG, H., BULLITT, E. and MARRON, J. (2009). A principal component analysis for trees. *The Annals of Applied Statistics* **1597**–1615.
- BARDEN, D., LE, H. and OWEN, M. (2013). Central limit theorems for Frechet means in the space of phylogenetic trees. *Electron. J. Probab* **18** 1–25.
- BECKMANN, C. F., DELUCA, M., DEVLIN, J. T. and SMITH, S. M. (2005). Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360** 1001–1013.
- BHATIA, R. (1997). *Matrix Analysis*. Springer, New York.
- BHATIA, R. (2007). *Positive Definite Matrices*. Princeton University Press, Princeton.
- BHATTACHARYA, A. and BHATTACHARYA, R. (2012). *Nonparametric Inference on Manifolds with Applications to Shape Spaces*. Cambridge University Press, New York.
- BHATTACHARYA, R. and LIN, L. (2013). A Central Limit Theorem for Frechet Means. *arXiv preprint arXiv:1306.5806* –.
- BHATTACHARYA, R. and PATRANGENARU, V. (2003). Large Sample Theory of Intrinsic and Extrinsic Sample Means on Manifolds. I. *The Annals of Statistics* **31** 1–29.
- BHATTACHARYA, R. and PATRANGENARU, V. (2005). Large sample theory of intrinsic and extrinsic sample means on manifolds. II. *The Annals of Statistics* **33** 1225–1259.
- BICKEL, P. J. and LEVINA, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics* 2577–2604.
- BICKEL, P. J. and LEVINA, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics* 199–227.
- BILLERA, L. J., HOLMES, S. P. and VOGTMANN, K. (2001). Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics* **27** 733–767.
- BISWAL, B. B., MENNES, M., ZUO, X.-N., GOHEL, S. and KELLY, C. E. A. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences* **107** 4734–4739.
- BONNABEL, S. and SEPULCHRE, R. (2009). Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM Journal on Matrix Analysis and Applications* **31** 1055–1070.
- BOOKSTEIN, F. (1978). *The Measurement of Biological Shape and Shape change*. Springer, London.
- BULLMORE, E. and SPORNS, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience* **10**(1) 1–13.
- CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* **106** 672–684.
- CAI, T., LIU, W. and LUO, X. (2011). A constrained  $L_1$  minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106** 594–607.
- CHAVEL, I. (1984). *Eigenvalues in Riemannian geometry*. Pure and Applied

- Mathematics* **115**. Academic Press, Inc., Orlando, FL.
- CHENG, S. H. and HIGHAM, N. J. (1998). A modified Cholesky algorithm based on a symmetric indefinite factorization. *SIAM Journal on Matrix Analysis and Applications* **19** 1097–1110.
- CHUNG, F. R. K. (1997). *Spectral graph theory* **92**. American mathematical society.
- DOUGHERTY, J., KOHAVI, R. and SAHAMI, M. (1995). Supervised and unsupervised discretization of continuous features. In *ICML* 194–202.
- DRYDEN, I. L., KOLOYDENKO, A. and ZHOU, D. (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Annals of Applied Statistics* **3** 1102–1123.
- ELLEGREN, H. and PARSCH, J. (2007). The evolution of sex-biased genes and sex-biased gene expression. *Nature Reviews Genetics* **8** 689–698.
- FISHER, R. (1953). Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **217** 295–305.
- FISHER, N. I., LEWIS, T. and EMBLETON, B. J. J. (1987). *Statistical analysis of spherical data*. Cambridge University Press.
- FRÉCHET, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de L’Institut Henri Poincaré* **10(4)** 215–310.
- FU, Y. and MA, Y. (2013). *Graph Embedding for Pattern Analysis*. Springer.
- GINESTET, C. E., FOURNEL, A. P. and SIMMONS, A. (2014). Statistical network analysis for functional MRI: Summary networks and group comparisons. *Frontiers in computational neuroscience* **8(51)** 1–10.
- GINESTET, C. E. and SIMMONS, A. (2011). Statistical Parametric Network Analysis of Functional Connectivity Dynamics during a Working Memory Task. *NeuroImage* **5(2)** 688–704.
- GROMOV, M. (2001). *Metric Structures for Riemannian and Non-Riemannian Spaces. Modern Birkhäuser Classics*. Birkhäuser, Berlin.
- HIGHAM, N. J. (2002). Computing the nearest correlation matrix: A problem from finance. *IMA Journal of Numerical Analysis* **22** 329–343.
- JAKOBSEN, S. K. (2014). Mutual information matrices are not always positive semi-definite. *IEEE Transactions on information theory* **60(5)** 0018.
- KANG, H., OMBAO, H., LINKLETTER, C., LONG, N. and BADRE, D. (2012). Spatio-spectral mixed-effects model for functional magnetic resonance imaging data. *Journal of the American Statistical Association* **107** 568–577.
- KAROUI, N. E. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics* 2717–2756.
- KENDALL, D. G. (1977). The diffusion of shape. *Advances in applied probability* 428–430.
- KENDALL, D. G. (1984). Shape manifolds, procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society* **16** 81–121.
- KENDALL, W. S. and LE, H. (2011). Limit theorems for empirical Fréchet means of independent and non-identically distributed manifold-valued random variables. *Brazilian Journal of Probability and Statistics* **25** 323–352.
- KRISHNAMACHARI, R. and VARANASI, M. (2013). On the geometry and quantization of manifolds of positive semi-definite matrices. *IEEE Transactions on*

- signal processing* **61** 4587 – 4599.
- LE, H. (2001). Locating Fréchet means with application to shape spaces. *Advances in Applied Probability* **33** 324–338.
- LE, H. and KUME, A. (2000). The Fréchet mean shape and the shape of the means. *Advances in Applied Probability* **32** 101–113.
- LEE, J. (2006). *Introduction to Smooth Manifolds*. Springer, London.
- LEE, H., LEE, D. S., KANG, H., KIM, B.-N. and CHUNG, M. K. (2011). Sparse brain network recovery under compressed sensing. *IEEE Transactions on Medical Imaging* **30** 1154–1165.
- LEON, P. S., KNOCK, S. A., WOODMAN, M. M., DOMIDE, L., MERSMANN, J., MCINTOSH, A. R. and JIRSA, V. (2013). The Virtual Brain: a simulator of primate brain network dynamics. *Frontiers in neuroinformatics* **7**.
- LINIAL, N. (2002). Finite metric spaces: combinatorics, geometry and algorithms. *Proceedings of the eighteenth annual symposium on Computational geometry* 63–63.
- LINIAL, N., LONDON, E. and RABINOVICH, Y. (1995). The geometry of graphs and some of its algorithmic applications. *Combinatorica* **15** 215–245.
- MARDIA, K. V. and JUPP, P. E. (2009). *Directional statistics* **494**. John Wiley & Sons.
- MCEWEN, B. S. (1999). Permanence of brain sex differences and structural plasticity of the adult brain. *Proceedings of the National Academy of Sciences* **96** 7128–7130.
- MICHELOYANNIS, S., VOURKAS, M., TSIRKA, V., KARAKONSTANTAKI, E., KANATSOULI, K. and STAM, C. J. (2009). The influence of ageing on complex brain networks: A graph theoretical analysis. *Human Brain Mapping* **30** 200–208.
- MOAKHER, M. (2005). A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications* **26** 735–747.
- MOAKHER, M. and ZERAI, M. (2011). The Riemannian Geometry of the Space of Positive-Definite Matrices and Its Application to the Regularization of Positive-Definite Matrix-Valued Data. *Journal of Mathematical Imaging and Vision* **40** 171–187.
- NEWMAN, M. (2010). *Networks: An Introduction*. Oxford University Press.
- PACHOU, E., VOURKAS, M., SIMOS, P., SMIT, D., STAM, C., TSIRKA, V. and MICHELOYANNIS, S. (2008). Working Memory in Schizophrenia: An EEG Study Using Power Spectrum and Coherence Analysis to Estimate Cortical Activation and Network Behavior. *Brain Topography* **21** 128–137.
- THIRION, B., FLANDIN, G., PINEL, P., ROCHE, A., CIUCIU, P. and POLINE, J.-B. (2006). Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets. *Hum. Brain Mapp.* **27** 678–693.
- TOMASI, D. and VOLKOW, N. D. (2010). Functional connectivity density mapping. *Proceedings of the National Academy of Sciences* **107** 9885–9890.
- TOMASI, D. and VOLKOW, N. D. (2011). Gender differences in brain functional connectivity density. *Human Brain Mapping* **33** 849–860.
- TZOURIO-MAZOYER, N., LANDEAU, B., PAPATHANASSIOU, D., CRIVELLO, F.,

- ETARD, O., DELCROIX, N., MAZOYER, B. and JOLIOT, M. (2002). Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain. *NeuroImage* **15** 273–289.
- WANG, H. and MARRON, J. (2007). Object oriented data analysis: Sets of trees. *The Annals of Statistics* **35** 1849–1873.
- WATSON, G. S. (1983). *Statistics on spheres* **6**. Wiley New York.
- WATTS, D. J. and STROGATZ, S. H. (1998). Collective dynamics of ‘small-world’ networks. *Nature* **393** 440–442.
- XIA, C. (2013). *Eigenvalues in Riemannian Geometry*. IMPA Mathematical Publications.
- YAN, S., XU, D., ZHANG, B., ZHANG, H.-J., YANG, Q. and LIN, S. (2007). Graph embedding and extensions: a general framework for dimensionality reduction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **29** 40–51.
- YAN, C.-G., CRADDOCK, R. C., ZUO, X.-N., ZANG, Y.-F. and MILHAM, M. P. (2013). Standardizing the intrinsic brain: towards robust measurement of inter-individual variation in 1000 functional connectomes. *Neuroimage* **80** 246–262.
- ZIEZOLD, H. (1977). On expected figures and a strong law of large numbers for random elements in quasi-metric spaces. *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 1974 European Meeting of Statisticians*.
- ZUO, X.-N., EHMKE, R., MENNES, M., IMPERATI, D., CASTELLANOS, F. X., SPORNS, O. and MILHAM, M. P. (2012). Network centrality in the human functional connectome. *Cerebral Cortex* **22** 1862–1875.